Final Report

1. Title of Project: Testing DNA samples for population of origin

2. Principal Investigator: Raymond D. Miller, Washington University

3. Reporting Period: Final Report, April. 1, 2007-April 30, 2008

4. Abstract:

To help field investigators with unidentified DNA samples, we chose 16 suitable genetic markers (single nucleotide polymorphisms), developed the theoretical basis of a forensic test to give information about the ethnic/population origin of a DNA sample, developed a practical implementation of the test using TaqMan genotyping technology, and began to implement the test in the crime laboratories of our collaborators. For a DNA sample, the genotyping results can be statistically compared with known population frequencies of self-described ethnic groups in order to narrow the possible ancestry of the sample.

5. Project Description:

Our proposal was responsive to the MFRC topic "tools that can reveal additional or more discriminatory information about forensic evidence," and was relevant to DNA analysis. Our collaborators have been the Missouri State Highway Patrol Crime Laboratory, St. Louis County Crime Laboratory, and the St. Charles Crime Laboratory.

To help field investigators with unidentified DNA samples, we proposed to develop and implement an efficient forensic test (we refer to as the "population test" or "ethnicity test") in the crime laboratories of our collaborators that would give information about the population of origin of the donor of a DNA sample. The test utilizes a highly selected group of single nucleotide polymorphisms (SNPs) that have very divergent allele frequencies between sample populations and are called ancestry informative markers (AIMs).

We utilized a number of criteria to filter a large set of AIMs to arrive at the 16 AIMs for the test. These were then extensively tested by *in silico* simulations to investigate the limits of the test. The original underlying hypothesis of this project was that a set of 16 AIMs could be chosen that would be very informative, and that hypothesis has been shown to be true.

The wet-lab portion of the test we developed was to genotype DNA samples using TaqMan genotyping assays. The equipment for genotyping was already present in the labs of our collaborators, and we provided them with reagents for the test. As a part of the test, the experimental results were compared with known population frequencies of self-described ethnic groups in order to narrow the possible ancestry of the sample.

Successful completion of this project has begun to implement a new forensic technique in the laboratories of our collaborators that provides information about the population of origin of a tested DNA sample for their field investigators. To expand usage of the test, the collaborators will help publicize the technique within the DNA forensic community.

6. Project Objectives:

The specific aims were the following: 1). Identify a subset of AIMs that will be useful for and sufficiently powerful for the proposed population test. 2). Assemble components of the population test, including chemistry and analysis tools, and validate them by characterizing

additional sets of DNAs. 3). Assist collaborators in implementing the population test, including providing components, protocols, and technical advice. 4). Publicize the population test to the wider forensics community.

7-8. Procedures, Results, and Discussion:

The specific aims of the project were broken down into steps as follows.

Aim 1. Identify a subset of AIMs that will be useful for and sufficiently powerful for the proposed population test.

Step 1a. Choose a set of 16 AIMs to be used in the ethnicity/population test. The AIMs were chosen from 165 SNPs characterized for allele frequencies in nine populations by Yang, N. *et al*, 2005, *Hum. Genet*. **118**:382-92. Using UNIX-based operating systems, we downloaded files containing these SNPs and allele frequencies. We used PERL scripts to filter this collection and compare with other databases, requiring that they could be matched with NCBI/dbSNP rs numbers, TaqMan assays were available from Applied BioSystems, and that the AIMs had been assayed in the HapMap project. We also removed AIMs from consideration that mapped to the X chromosome because they reduced the power of the test and increased the complexity of the computations. We removed G/C and A/T AIMs, since in our experience these frequently lead to confusion because the same alleles are present in both DNA orientations. Then we selected by hand 16 AIMs such that each is highly divergent in frequency between two or more populations of different continental origin, and all are widely separated in the genome. The set of AIMs is given on our website (http://snp.wustl.edu).

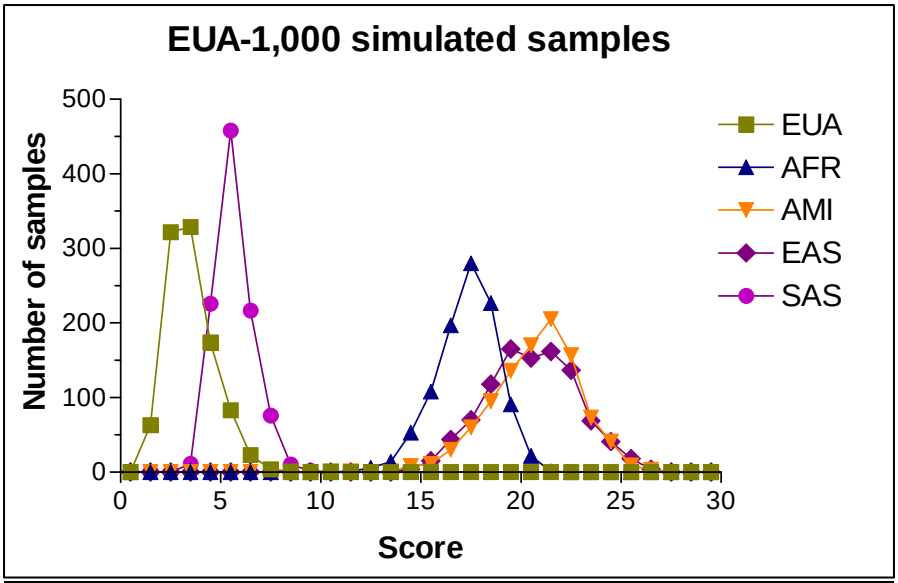

Figure 1. Population test results for 1,000 simulated European-derived samples.

Step 1b. Complete the theoretical framework for the population test. The method is based upon conditional probabilities using the allele frequencies within the populations studied by Yang *et al*. For the 16-AIM genotype of an individual, the calculated probability of observing that genotype is much lower if frequencies from an incorrect population are used than if those from the correct population are used. Therefore, many populations can be ruled out as the source. We compute a score for an individual based on each population. To give a convenient range, we use the negative sum of the log10 values of each probability. We have constructed an Excel spreadsheet that calculates the answers.

Step 1c. Perform extensive *in silico* work to test our selection of AIMs and the theoretical framework for the population test. Based upon the known population frequencies, we used a computer random number generator and PERL scripts to simulate the 16-AIM genotypes for 9,000 individuals, 1,000 from each of the nine populations. We computed test and summary statistics for this dataset. Then we repeated the entire exercise to check consistency. For example, each of 1,000 simulated individuals from a European-derived population (EUA) can be clearly excluded from African (AFR), East Asian (EAS), and Amerind (AMI) populations (Fig. 1) and from an African American population (not shown). However, they cannot always be excluded from the South Asian (SAS) population (Fig. 1) or Puerto Rican population (not shown). For a particular sample, the test does not tell which population the sample is from, but it can exclude many populations—particularly those of different continental origin—as the source of the sample.

Step 1d. Obtain genotypes for our chosen 16 AIMs from humans of known population origin, and use the genotypes to further test the theoretical framework of the population test. We mined the HapMap Project database to obtain genotypes for the selected AIMs in 90 CEPH trios of European descent (CEU), 90 Yoruba trios from Nigeria (YRI), 45 unrelated Han Chinese from Beijing (CHB), and 45 unrelated Japanese from Tokyo (JPT). For each of the individuals except one, populations of origin on different continents were excluded. For the exception, many of the genotypes were missing, demonstrating one complexity of actual data sets. With missing data, the population test is conservative, possibly failing to exclude some populations.

Aim 2. Assemble components of the population test, including chemistry and analysis tools, and validate them by characterizing additional sets of DNAs.

Step 2a. Order the TaqMan reagents for the test. The assays were ordered from Applied Biosystems as “medium syntheses,” larger, more cost efficient quantities, and then subdivided for our use and for our collaborators. Other assay components were also ordered in larger quantities and subdivided. We verified that we could robustly do the test in 10 micro liter volumes with the Applied Biosystems 7000 real-time PCR machine. All assays were successfully tested using the 96-well CEU HapMap DNA sample collection. Although the HapMap used different genotyping technologies, we verified that we scored the same genotypes as they did, as expected, and we filled in some data missing in the HapMap project.

Step 2b. Make the TaqMan test more efficient. The TaqMan software is designed to run and analyze a few SNPs for many DNAs, e.g. one SNP for a plate of 96 DNA samples. For many forensic situations, the lab will have only a few samples, and for maximal power, they will need to analyze the samples for 16 SNPs. In other words instead of analyzing a few SNPs for many DNAs, it would be much more efficient to be able to analyze many (16) SNPs for a few DNAs. The basic output from the Applied Biosystems TaqMan genotyping run is an X and a Y value for each sample corresponding to the final fluorescence of each of the two dyes in the reaction. The usual software produces an X-Y plot of these values, and then has some clustering tools to distinguish the four groups: two homozygotes, the heterozygote, and the group with no incorporation (blanks). We observed that analysis of each of the SNPs produced good clusters that were reproducible in duplicate runs. However, the exact location of the clusters differed among the SNPs. We reasoned these differences were due to slight assay-specific and allele-

specific variation in efficiencies of the chemical reactions. We estimated parameters for each of the assays to normalize the assays to a common plot.

Step 2c. Produce software for the analysis of the population test. We constructed Excel spreadsheets to allow TaqMan results for the 16 AIMs and 1-6 DNAs to be efficiently analyzed. We also have constructed an Excel spreadsheet to convert a 16-AIM genotype to population scores and showing excluded populations.

Aim 3. Assist collaborators in implementing the population test, including providing components, protocols, and technical advice.

Step 3a. Meetings. We introduced our collaborators to the population test in a two hour meeting at Washington University School of Medicine, on March 19, 2008. In attendance were two people from Missouri Highway Patrol Crime Lab, three people from the St. Charles County Crime lab, and one person from the St. Louis County Crime lab.

Step 3b. Other assistance. We have had numerous phone calls and emails and a few visits with our collaborators. The time within the grant ran short, and the assistance process is still in progress. More interaction will be required to have the population test fully operational in the collaborator's labs.

Aim 4. Publicize the population test to the wider forensics community.

Step 4a. Website. http://snp.wustl.edu/

Step 4b. Publications. This work was briefly described in a news and opinion feature article (Melissa Lee Phillips, *BioScience* **58**: 480, June 2008).

A major publication on this work is in preparation.

Step 4c. Meetings. We anticipate that our collaborators will present their results at some meetings in the future.

9. Dissemination Discussion:

See Aim 4 above.

10. Appendix.

a. Discussion of problems that have arisen.

None, however, there is a need for more training for our collaborators.